Reducing semantic complexity in distributed digital libraries: treatment of term vagueness and document re-ranking

Mayr, Philipp; Mutschke, Peter; Petras, Vivien

Empfohlene Zitierung / Suggested Citation:
Mayr, P., Mutschke, P., & Petras, V. (2008). Reducing semantic complexity in distributed digital libraries: treatment of term vagueness and document re-ranking. *Library Review*, *57*(3), 213-224. https://doi.org/10.1108/00242530810865484

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0024-2535.htm

Reducing semantic complexity in distributed digital libraries

Treatment of term vagueness and document re-ranking

Philipp Mayr, Peter Mutschke and Vivien Petras
GESIS-IZ Social Science Information Centre, Bonn, Germany

Received 19 October 2007
Reviewed 9 October 2007
Accepted 13 November 2007

Abstract

Purpose – The general science portal "vascoda" merges structured, high-quality information collections from more than 40 providers on the basis of search engine technology (FAST) and a concept which treats semantic heterogeneity between different controlled vocabularies. First experiences with the portal show some weaknesses of this approach which come out in most metadata-driven Digital Libraries (DLs) or subject specific portals. The purpose of the paper is to propose models to reduce the semantic complexity in heterogeneous DLs. The aim is to introduce value-added services (treatment of term vagueness and document re-ranking) that gain a certain quality in DLs if they are combined with heterogeneity components established in the project "Competence Center Modeling and Treatment of Semantic Heterogeneity".

Design/methodology/approach – Two methods, which are derived from scientometrics and network analysis, will be implemented with the objective to re-rank result sets by the following structural properties: the ranking of the results by core journals (so-called Bradfordizing) and ranking by centrality of authors in co-authorship networks.

Findings – The methods, which will be implemented, focus on the query and on the result side of a search and are designed to positively influence each other. Conceptually, they will improve the search quality and guarantee that the most relevant documents in result sets will be ranked higher.

Originality/value – The central impact of the paper focuses on the integration of three structural value-adding methods, which aim at reducing the semantic complexity represented in distributed DLs at several stages in the information retrieval process: query construction, search and ranking and re-ranking.

Keywords Digital libraries, Worldwide web, Information management

Paper type Research paper

Introduction

In the area of scientific and academic information systems, a whole array of bibliographic databases, disciplinary Internet portals, institutional repositories or archival and other media type collections are increasingly accumulated and embedded in all-encompassing information systems. Such collections are necessary in order to meet user expectations that demand one-stop "information fulfillment". Examples are Elsevier's Scirus portal[1], the Online Computer Library Center WorldCat union catalog[2] or Tuft University's Perseus project[3].

In Germany, an ambitious project for one-stop academic search is the vascoda portal[4], a joint project between the BMBF (Federal Ministry for Education and Research) and the DFG (German Research Foundation). Vascoda provides a federated search interface for a multitude of disciplinary and interdisciplinary databases (e.g. full-text article databases, indexing and abstracting services, library catalogs) and internet resource collections.

The vascoda portal contains many information collections that are meticulously developed and structured. They have sophisticated subject metadata schemes (subject headings, thesauri or classifications) to describe and organise the content of the

Library Review
Vol. 57 No. 3, 2008
pp. 213-224

0024-2535
DOI 10.1108/00242530810865484

documents on an individual collection level. The general search interface, however, only provides a free-text search over all metadata fields without regard for the precise subject access tools that were originally intended for these information collections.

If large-scale contemporary information organisation efforts like the Semantic Web[5] (see also Krause, 2006, 2007, 2008) strive to provide more structure and semantic resolution with respect to information content, how is it possible that advanced interfaces for digital libraries (DLs) scale back on exactly the same issue?.

Search – both in full-text collections like the Internet or more heavily structured and less diverse collections like institutional repositories, indexing databases or library catalogs as described above – only works as well as the matching between the language in queries and the language in the searched documents. If the words in the query are different from the words in a relevant document, this document will not be found. The problem of matching query terms to document terms is a result of the ambiguity or vagueness of language (Blair, 1990, 2003).

Because of the sheer size and variation of large full-text databases, this problem is not as noticeable because any query (even if they contain spelling mistakes or nonsense statements) will find documents. The problem is aggravated in collections of more restricted volume or text (i.e. repositories that contain only formal metadata, some subject description and just a link to the full-text). The issue becomes even more critical when several collections with different metadata schemes are searched at the same time – which is the case in the distributed search scenario. In this scenario, not only is the matching between query and document terms affected by language ambiguity, but also the matching between different subject-describing metadata schemes. In Figure 1, we speak of vagueness 1 and vagueness 2/3 (V1 and V2/3) to denote the different areas where language ambiguity can occur. For successful retrieval in any DL, both levels of vagueness have to be addressed (compare Hellweg *et al.*, 2001).

Furthermore, the result sets of transformed or expanded queries in distributed collections are often very large and tests show that the conventional web-based ranking methods are not appropriate for the heterogeneous metadata records. Therefore, two methods, which are derived from scientometrics and network analysis, will be implemented with the objective to re-rank result sets: (a) the ranking of the results by core journals (so-called Bradfordizing) and (b) ranking by centrality of authors in co-authorship networks.

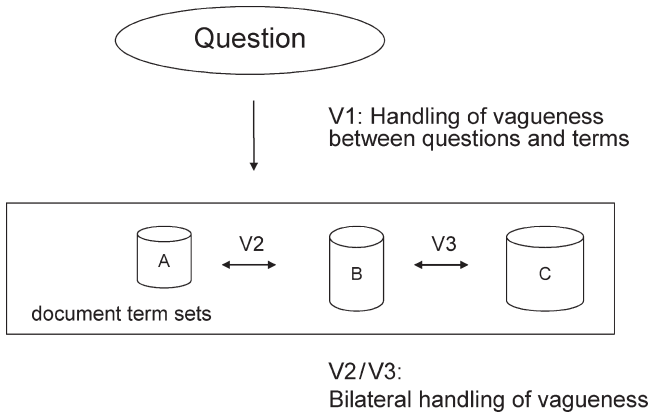

Figure 1.
Two step methodology of vagueness treatment

This paper is a description of an attempt to harness the semantic knowledge in controlled vocabularies for several stages in the information retrieval process: query construction, search and ranking and re-ranking. We briefly describe the GESIS project "Competence Center Modeling and Treatment of Semantic Heterogeneity" with the goal of creating a semantic network of terms in different controlled vocabularies (terminology mapping) in order to facilitate a seamless search across different subject-based knowledge organisation systems. At the conclusion of this project, we will discuss modules that are being devised to leverage these mappings for an improved user search experience.

Results from a major terminology mapping effort

Semantic integration seeks to connect different information systems through their subject metadata frameworks; insuring that distributed searches over several information systems can still use the advanced subject access tools provided with the individual databases. Through the mapping of different subject terminologies, a "semantic agreement" for the overall collection to be searched is achieved. Terminology mapping – the mapping of words and phrases of one controlled vocabulary to the words and phrases of another – creates a semantic network between the information systems carrying the advantages of controlled subject metadata schemes into the distributed DL world.

In 2004, the German Federal Ministry for Education and Research funded a major terminology mapping initiative at the GESIS Social Science Information Centre in Bonn (GESIS-IZ) "Competence Center Modeling and Treatment of Semantic Heterogeneity"[6], which concluded this year (see Mayr/Walter, 2007a, b). The task of this terminology mapping initiative was to organise, create and manage "cross-concordances" between major controlled vocabularies (thesauri, classification systems, subject heading lists), centred around the social sciences but quickly extending to other subject areas (e.g. political science, economics, medicine or subject-specific parts of universal vocabularies). Cross-concordances are intellectually (manually) created crosswalks that determine equivalence, hierarchy and association relations between terms from two controlled vocabularies. Most vocabularies in the project have been related bilaterally; that is, there is a cross-concordance relating terms from vocabulary A to vocabulary B as well as a cross-concordance relating terms from vocabulary B to vocabulary A (note: bilateral relations are not necessarily symmetrical). Other definitions and examples of crosswalks between controlled vocabularies exist in an international context (see overview in Zeng/Chan, 2004; Vizine-Goetz *et al.*, 2004; Liang and Sini, 2006).

In November 2007, 25 controlled vocabularies from 11 disciplines were connected with vocabulary sizes ranging from 1,000 to 17,000 terms per vocabulary. To date, more than 513,000 relations in 64 crosswalks have been generated. An overview of the preliminary project results presented at the NKOS/ECDL workshop 2007 can be found in[7].

A database including all mapped controlled terms and cross-concordance relations was built and a "heterogeneity service" developed. The heterogeneity service is a web service, which makes the cross-concordances available for other applications (see Figure 2). Many cross-concordances are already implemented and utilised for the German Social Science Information Portal sowiport[8], which searches bibliographical and other information resources (including 13 databases with 10 different vocabularies and about 2.5 million references).

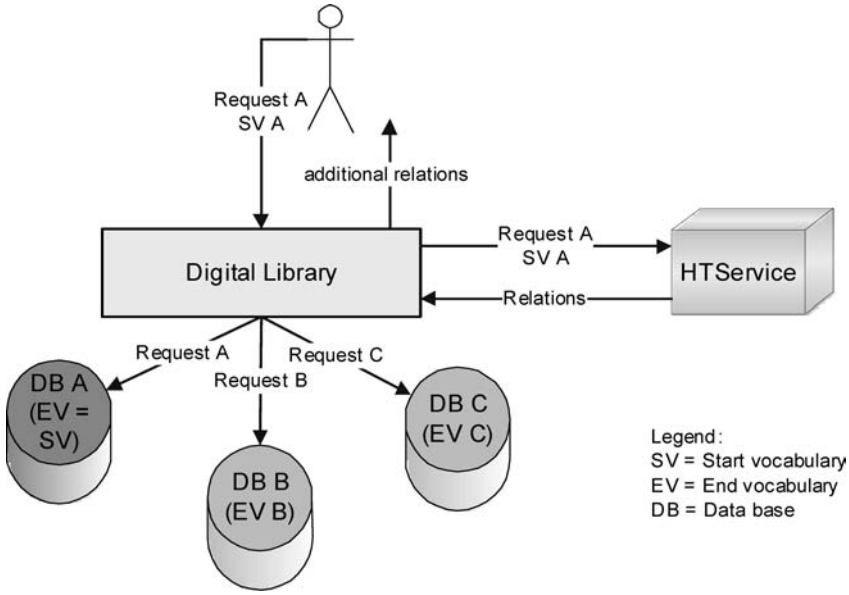

Figure 2.
Heterogeneity service (HTS)

Semantic mappings could support distributed searching in several ways. First and foremost, they should enable seamless searching in databases with different subject metadata systems. Additionally, they can serve as tools for vocabulary expansion since they present a vocabulary network of equivalent, broader, narrower and related term relationships. Thirdly, this vocabulary network of semantic mappings can also be used for query expansion and reformulation.

The following section introduces the concept of a search term recommender (STR). This tool is an aid for query reformulation and reconstruction which has been adapted for a web-based information portal from human search intermediaries (e.g. reference librarians).

Search term recommender

Semantic mappings can reduce the problem of language ambiguity at the vagueness 2/3 layer described in Figure 1 (between different information systems). However, the vagueness at the user-information system interface remains unaddressed.

To reduce language ambiguity at the vagueness 1 layer (between query terms and document terms), another instrument is necessary to map terms at this interface. The goal of this mapping would be to "translate" the query terms of a user to the document terms of the database (or vice versa) in order to produce a match at search time. Since we are mostly concerned with information systems that contain sparse text enriched with subject describing controlled vocabularies, we propose a STR system which will propose terms from the controlled vocabularies, given a specified query.

The basic parameters of a search term suggestion system are the controlled vocabulary terms that are used for document representation and the natural language keywords that are input by the searcher. The advantage of suggesting controlled vocabulary terms as search terms is that these terms have been systematically assigned to the documents, so that there is a high probability of relevant and precise

retrieval results if these terms are used instead of whatever natural language keywords the searcher happens to think of.

A second advantage in suggesting controlled vocabulary terms is their application in the semantic network of the cross-concordances. That is, if controlled vocabulary terms are used in searching, the cross-concordances, which map these terms between different databases, can be successfully applied for distributed retrieval.

In addition, this kind of vocabulary help will hopefully improve the search experience for the user in general. Suggesting terms reduces the searcher's need to think of other relevant search terms that might describe his or her information need. It effectively eases the cognitive load on the searcher since it is much easier for a person to pick appropriate search terms from a list than to come up with search terms by themselves. It also helps to alleviate "anchoring bias" (Blair, 2002), which is an effect that makes it harder to substantially deviate from one's original thought-of search terms and to consider different search terms or strategies.

Another consequence of term suggestion is the presentation of new or different technical expressions for a concept. This again could lead to changes in a search strategy or topic, which might help in reaching the user search goal. Term suggestions from several fields of research and/or information resources could also provide an overview over different areas of discussion, which deal with particular concepts (perhaps assuming different meanings or directions of thought). The result would be a different domain perspective on certain concepts, an effect which can also be achieved by displaying the semantic mappings of the cross-concordances themselves.

A STR is created by building a dictionary of associations between two vocabularies:

(1) natural language terms and phrases from the documents in the information collection (e.g. titles, abstracts, authors) and,

(2) the controlled vocabulary (thesaurus terms, subject headings, classification numbers, etc.) used for document representation.

In one implementation, a likelihood ratio statistic is used to measure the association between the natural language terms from the collection and the controlled vocabulary terms to predict which of the controlled vocabulary terms best mirror the topic represented by the searcher's search terms (Plaunt and Norgard, 1998; Gey *et al.*, 1999). However, other methods of associating natural language terms and controlled vocabulary terms are possible (Larson, 1991, 1992).

In an information system with several information resources (i.e. databases) and several controlled vocabularies, a search term recommendation tool has to determine which terms from which vocabularies to suggest to the user and how to tie the term suggestions for query construction into the semantic network of cross-concordances. Several approaches seem possible: a pivot controlled vocabulary, from which terms are suggested and mappings approached; a general suggestion pattern, which clusters similar concepts from several vocabularies; or a domain-specific approach, whereby terms and vocabularies are chosen according to the subject of interest for the searcher.

The result sets of transformed or expanded queries in distributed collections are often very large and tests show that the conventional web-based ranking methods are not appropriate for presenting heterogeneous metadata records as suitable result sets to the user. In the following section, we propose re-ranking methods (implemented as post-search modules), which are based on structures and regularities in scientometrics and network analysis.

Re-ranking

Compared to traditional text-oriented sorting mechanisms, our scientometric and network analysis re-ranking methods offer a completely new view on results sets, which have not been implemented in heterogeneous and larger database scenarios to date. The usage of these modules should be an alternative ranking opportunity with the objective to enhance and improve the search process in general. In addition, we expect an improvement in document relevance for the top-listed documents.

Bradford Law of Scattering and Bradfordizing

Bradford Law of Scattering and Bradfordizing have their roots in scientometrics and are often applied in bibliometric analyses of databases and collections as a tool for systematic collection management in library and information science. Fundamentally, Bradford Law states that literature on any scientific field or subject-specific topic scatters in a typical way. A core or nucleus with the highest concentration of papers (a few core journals) on a topic is followed by zones with loose concentrations of paper frequency, which is described by Bradford:

> . . . if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the numbers of periodicals in the nucleus and succeeding zones will be as $1{:}n{:}n^2$ . . . (Bradford, 1948)

Bradford Law as a general law in informetrics can be applied to all scientific disciplines, and especially in a multi-database scenario and in combination with the aforementioned semantic treatment of heterogeneity.

Bradfordizing (White, 1981) is an information science application of the Bradford Law of Scattering which sorts/re-ranks a result set according to the identified core journals for a query. The journals for a search are ranked by the frequency of their listing in the result set (number of articles for a journal title). If a search result is bradfordized, articles of core journals are ranked ahead of the journals which contain an average number or only few articles on a topic. This method is interesting in the context of our re-ranking task because it is a robust way of sorting the central publication sources for any query to the top positions of a result set. Bradfordizing has the following values-added:

(1) An alternative view on results sets which are ordered by core journals (the user is provided with documents of core journals first);

(2) An alternative view on publication sources in an information space which are intuitively closer at the research process than statistical methods (e.g. best match) or traditional methods (e.g. exact match);

(3) Possibly a higher topical relevance of re-ranked documents.

Additionally, re-ranking via bradfordized lists offer an opportunity to switch between term-based search and the alternative search mode browsing. Bates (2002) brings together Bradford Law and information seeking behavior.

> . . . the key point is that the distribution tells us that information is neither randomly scattered, nor handily concentrated in a single location. Instead, information scatters in a characteristic pattern, a pattern that should have obvious implications for how that information can most successfully and efficiently be sought (Bates, 2002).

Bates applies conceptually different search techniques (directed searching, browsing and linking) to the Bradford zones. Bates postulates the Bradford nucleus for browsing, the second zone for directed searching with search terms and further zones for linking. We focus on an automatic change from directed searching (enhanced by treatment of semantic heterogeneity) into browsing. Starting with a subject specific descriptor search, we will connect the query with our heterogeneity service to transfer descriptor terms into a multi-database scenario. In the second step, the results from the different databases will be combined and sorted according to Bradford's method (i.e. most productive journals for a topic first). The conclusion this step provides us with a Bradfordized list of journal articles. The next step is the extraction of a result set of all documents in the Bradford nucleus which can be delivered for browsing. This automatically generated browsing modus can be compared to Bates search technique "journal run".

The focus on a re-ranking technique based on Bradfordizing is interesting owing to the universal properties of the law, allowing it to be applied in a one-database scenario (e.g. Mayr and Umstätter, 2007) and a multi-database scenario like vascoda or sowiport. On a very abstract level, Bradford re-ranking can be used as a compensation method for enlarged search spaces; however, in our application model the information on the core journals is used for document ranking.

Co-author networks

It is generally acknowledged that standard search services do not meet the wealth of information material supplied by DLs. Traditional retrieval systems are strictly document oriented such that a user is unable to exploit the full complexity of the information stored therein. Bibliographic data, for instance, offers a rich information structure that is usually "hidden" in traditional information systems. A typical example of this issue are link structures among authors, given, for instance by co-author relationships, and – more importantly – the strategic position of authors within a given collaboration structure. Moreover, relevant information is more and more distributed over several heterogeneous information sources and services (e.g. bibliographic reference services, citation indices, full-text services, etc.).

DLs are therefore only meaningfully usable if they provide high-level search services that fully exhaust the information structures stored and, at the same time, reduce the complexity of information to highly relevant items. Due to the notion of a Semantic Web, particularly the Friend of a Friend approach[9], this strongly suggests the development of techniques that overcome the strict document orientation of standard indexing and retrieval methods by providing a deeper analysis of link structures and the centrality of entities in a given network structure.

This approach focuses on network analysis concepts for extracting central actors in co-author networks and ranking documents by author centrality (Mutschke, 2003). The expressiveness of co-author networks has been demonstrated in a number of scientometric studies (see e.g. Beaver, 2004). The basic approach of our model is to reason about the network structure in order to evaluate relevant authors for a particular domain. This information on the centrality of authors within their scientific community is then used to rank documents.

According to graph theory, a co-author network in our model is described as a graph $G = (V, E)$, where V is the set of vertices (authors), and E the set of edges (co-authorships). A co-author network is generated on the basis of all co-authorships that appear in a given document set (e.g. the result set of a query). On social networks a

number of calculations can be performed. An important structural attribute of the vertices is their centrality. Centrality measures the contribution of a network position to the vertices' prominence, influence or importance within a social structure. In our model, we use the betweenness measure. Betweenness focuses on the ratio of shortest paths a vertex lies on. An author with a high betweenness is therefore a vertex that connects many authors in the network. Betweenness is therefore seen as a measure indicating an actor's degree of control or influence of the interaction processes that construct a network structure.

Accordingly, an index of centrality within a scientific collaboration and communication structure might indicate the degree of relevance of an author for the domain in question; such relevance would be attributable to his/her key position in the network. In our application model information on the centrality of authors is used for document ranking. This is done by weighting the documents retrieved by the centrality values of their authors such that the user is provided with documents of central authors.

Figure 3 visualises the planned application of the value-added services (in the stages of the search process and the combination of the single components). See the STR in the beginning of a search and re-ranking of combined search result sets at the end of a search loop (see stages 2 and 7 in Figure 3).

Integration

Beyond an isolated use, a combination of the approaches is promising to yield much higher innovation potential. In our model, the following scenarios are supported (e.g. combining Bradfordizing with Author Centrality as in Figure 4).

The user is provided with publications which are associated with both central authors as well as core journals. From a technical point of view, the following variants are suitable and may yield different results:

- Bradfordizing as a filter for the network analysis process: central authors are evaluated within the set of documents which are associated with core journals (i.e. the result set is reduced to the core journal set before author centrality analysis is performed).
- Author centrality as a filter for Bradfordizing (the "inverse" version of the variant above): Bradfordizing is performed on the set of result set document which are assigned to central authors (i.e. the result set is reduced to "central" documents before core journals are evaluated).
- The "intersection" variant: core journals and central authors are first evaluated independently from one another on the basis of the whole result set. Publications that satisfy both relevance criteria (they appear in a core journal and their authors are central) are determined in a second step (see Figure 4).

Those combination models could not only be applied to result set re-ranking but also to the search term recommendation process (i.e. the usage of Bradfordizing and author centrality analysis as a filter on the collection used for the STR analysis).

Future research work should address the use of information pertaining to institutions, themes or citations as a means of providng further value-adding functions in re-ranking methods, rather than just using authors and journals. An important further research issue is to apply and evaluate the proposed ranking methods at the user search stage in order to improve the precision of the initial result set.

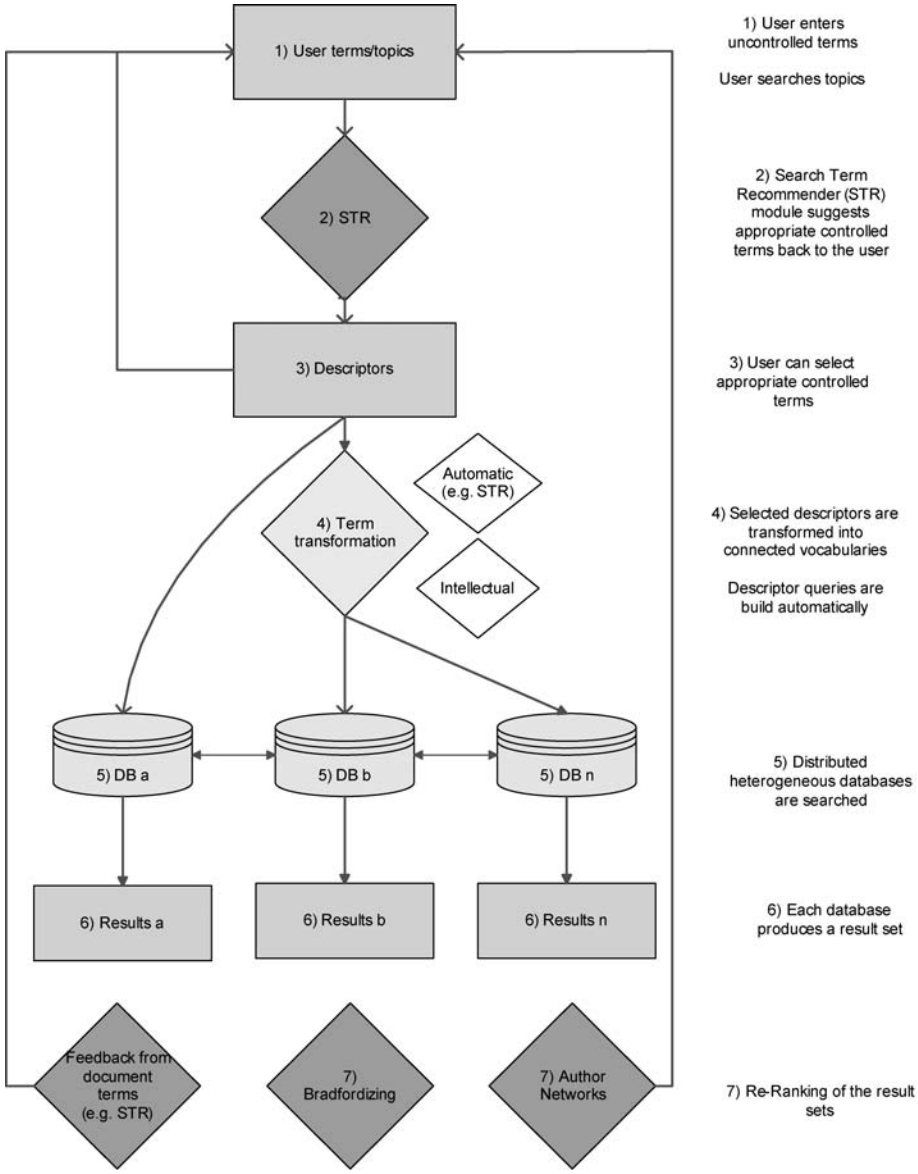

Figure 3.
Combination and embedding of the modules: STR and re-ranking

Outlook

The central impact of the paper focuses on the integration of three structural value-adding methods which aim at reducing the semantic complexity represented in distributed DLs at several stages in the information retrieval process: query construction, search and ranking and re-ranking. The integration of the models will be done using Semantic Web technologies which should enhance further insights into the usage of these techniques. The intersection of the Semantic Web world with the DL world as mentioned in Krause (2008) (available in this issue of Library Review) will

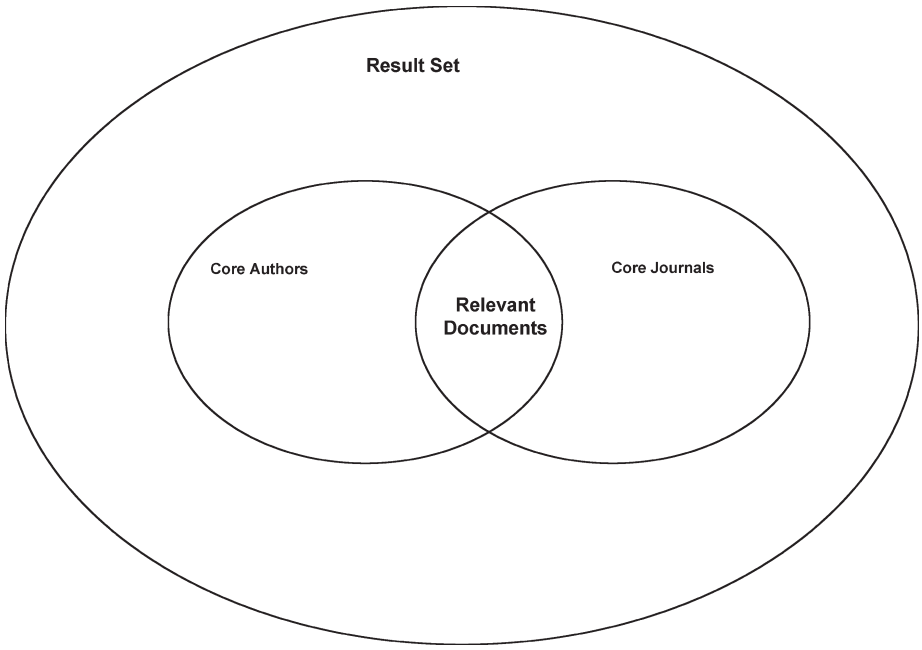

Figure 4.
Intersection of core journal and central author documents

hopefully result in more sophisticated analytical tools and interfaces for the presentation of information adapted to users' needs.

Notes

1. Elsevier. "Scirus – for scientific information only". Retrieved October 2007, from www.scirus.com/.
2. Online Computer Library Center (OCLC). "WorldCat". Retrieved October 2007, from www.oclc.org/worldcat/.
3. Department of the Classics, Tufts University. "The Perseus Digital Library". Retrieved October 2007, from http://perseus.mpiwg-berlin.mpg.de/.
4. "vascoda - Entdecke Information". Retrieved October 2007, from www.vascoda.de/.
5. World Wide Web Consortium (W3C). (2001). "Semantic Web Activity". Retrieved October 2007, from www.w3.org/2001/sw/.
6. "Competence Center Modeling and Treatment of Semantic Heterogeneity". Retrieved October 2007, from www.gesis.org/en/research/information_technology/komohe.htm.
7. Results from a German terminology mapping effort: intra- and interdisciplinary cross-concordances between controlled vocabularies. Presented at the NKOS/ECDL Workshop in Budapest Hungary. Retrieved October 2007, from http://dlist.sir.arizona.edu/2054/.
8. Sowiport. Retrieved October 2007, from www.sowiport.de.
9. Friend-of-a-Friend (FOAF). Retrieved October 2007, from http://xmlns.com/foaf/spec/.
10. For this variant a (configurable) threshold for centrality is needed.

References

Bates, M.J. (2002), "Speculations on browsing, directed searching, and linking in relation to the bradford distribution", in Bruce, H., Fidel, R., Ingwersen, P. and Vakkari, P. (Eds.), *Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)*, available at: www.gseis.ucla.edu/faculty/bates/articles/Searching_Bradford-m020430.html (accessed 20 December 2007).

Beaver, D. (2004), "Does collaborative research have greater epistemic authority?", *Scientometrics*, Vol. 60 No. 3, pp. 309-408.

Blair, D.C. (1990), *Language and Representation in Information Retrieval*, Elsevier Science Publishers Amsterdam, New York, NY, p. 335.

Blair, D.C. (2002), The challenge of commercial document retrieval, part II: a strategy for document searching based on identifiable document partitions", *Information Processing and Management*, Vol. 38 No. 2, pp. 293-304.

Blair, D.C. (2003), "Information retrieval and the philosophy of language", *Annual Review of Information Science and Technology*, Vol. 37, pp. 3-50.

Bradford, S.C. (1948), *Documentation*, Lockwood, London, p. 156.

Gey, F., Chen, H., Norgard, B., Buckland, M., Kim, Y., Chen, A., Lam, B., Purat, Y. and Larson, R. (1999), "Advanced search technology for unfamiliar metadata", *Third IEEE Metadata Conference*, Bethesda, MD.

Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M.N.O., Mutschke, P. and Strötgen, R. (2001), "Treatment of semantic heterogeneity in information retrieval", IZ Working Paper; No 23, IZ Sozialwissenschaften, Bonn, p. 47, available at: www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_23.pdf (accessed 20 December 2007).

Krause, J. (2006), "Shell model, semantic web and web information retrieval", in Harms, I., Luckhardt, H.-D. and Giessen, H.W. (Eds), *Information und Sprache: Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern, Festschrift für Harald H. Zimmermann*,Saur, München, pp. 95-106.

Krause, J. (2007), "The concepts of semantic heterogeneity and ontology of the Semantic Web as a background of the German science Portals vascoda and sowiport", in Prasad, A.R.D. and Madalli, D.P. (Eds), *International Conference on Semantic Web and Digital Libraries (ICSD 2007)*, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore, pp. 13-24, available at: https://drtc.isibang.ac.in/bitstream/1849/307/1/002_p39_krause_germany_formatted.pdf (accessed 20 December 2007).

Krause, J. (2008), "Semantic heterogeneity: comparing new Semantic Web approaches with those of digital libraries", *Library Review*, Vol. 57 No. 3.

Larson, R.R. (1991) "Classification clustering, probabilistic information-retrieval, and the online catalog", *Library Quarterly*, Vol. 61 No. 2, pp. 133-73.

Larson, R.R. (1992), "Experiments in automatic library-of-congress classification", *Journal of the American Society for Information Science*, Vol. 43 No. 2, pp. 130-48.

Liang, A.C. and Sini, M. (2006), "Mapping AGROVOC and the Chinese Agricultural Thesaurus: definitions, tools, procedures", *New Review in Hypermedia and Multimedia*, Vol. 12 No. 1, pp. 51-62.

Mayr, P. and Umstätter, W. (2007), "Why is a new *Journal of Informetrics* needed?", *Cybermetrics*, Vol. 11 No. 1, available at: www.cindoc.csic.es/cybermetrics/articles/v11i1p1.html (accessed 20 December 2007).

Mayr, P. and Walter, A.-K. (2007a), "Einsatzmöglichkeiten von Crosskonkordanzen", in Stempfhuber, M. (Ed.), *Lokal - Global: Vernetzung wissenschaftlicher Infrastrukturen: 12, Kongress der IuK-Initiative der Wissenschaftlichen Fachgesellschaft in Deutschland*, GESIS – IZ Sozialwissenschaften, Bonn, pp. 149-66, available at: www.gesis.org/Information/

Forschungsuebersichten/Tagungsberichte/Vernetzung/Mayr-Walter.pdf (accessed 20 December 2007).

Mayr, P. and Walter, A.-K. (2007b), "Zum Stand der Heterogenitätsbehandlung in vascoda: Bestandsaufnahme und Ausblick", in BID (Ed.), *Information und Ethik 3. Leipziger Kongress für Information und Bibliothek*, Verlag Dinges and Frick, Leipzig, available at: www.opus-bayern.de/bib-info/volltexte/2007/290/ (accessed 20 December 2007).

Mutschke, P. (2003), "Mining networks and central entities in digital libraries: a graph theoretic approach applied to co-author networks", *IDA 2003 – The fifth International Symposium on Intelligent Data Analysis*, *Berlin*, available at: http://fuzzy.cs.uni-magdeburg.de/confs/ida2003/ (accessed 20 December 2007).

Plaunt, C. and Norgard, B.A. (1998), "An association-based method for automatic indexing with a controlled vocabulary", *Journal of the American Society for Information Science*, Vol. 49 No. 10, pp. 888-902.

Vizine-Goetz, D., Hickey, C., Houghton, A. and Thompsen, R. (2004), "Vocabulary mapping for terminology services", *Journal of Digital Information*, Vol. 4 No. 4, available at: http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/ (accessed 20 December 2007).

White, H.D. (1981), ""Bradfordizing" search output: how it would help online users", *Online Review*, Vol. 5 No. 1, pp. 47-54.

Zeng, M.L. and Chan, L.M. (2004), "Trends and issues in establishing interoperability among knowledge organization systems", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 3, pp. 377-95.

Corresponding author
Philipp Mayr can be contacted at: philipp.mayr@gesis.org